

**BÁO CÁO HADOOPSTREMING**

**Môn học: BIG DATA**

**GVHD: TS. Lê Thị Minh Châu**

**SVTH: Trần Minh Đức**

**MSSV: 20133037**

**Thành Phố Hồ Chính Minh – Tháng 10 2022**

**1.Task1**

1. How do the input keys/values, the intermediate keys/values and the output keys/values relate?

Trong thực thi công việc MapReduce, trước khi gửi dữ liệu đến trình ánh xạ, trước tiên hãy chuyển đổi dữ liệu đó thành các cặp khóa-giá trị. Bởi vì ánh xạ chỉ các cặp dữ liệu khóa-giá trị.

* Cặp khóa-giá trị trong MapReduce được tạo như sau:

InputSplit - Nó là biểu diễn logic của dữ liệu mà InputFormat tạo ra. Trong chương trình MapReduce, nó mô tả một đơn vị công việc có chứa một nhiệm vụ bản đồ duy nhất.

RecordReader - Nó giao tiếp với InputSplit. Sau đó, nó chuyển đổi dữ liệu thành các cặp giá trị khóa phù hợp để Mapper đọc. RecordReader theo mặc định sử dụng TextInputFormat để chuyển đổi dữ liệu thành các cặp giá trị khóa.

Trong thực thi công việc MapReduce, hàm bản đồ xử lý một cặp khóa-giá trị nhất định. Sau đó, phát ra một số cặp khóa-giá trị nhất định. Hàm Reduce xử lý các giá trị được nhóm theo cùng một khóa.

1. How does MapReduce deal with node failures?

* Worker Failure:

- Master sẽ gửi tín hiểu tới các từng worker node còn lại.

- Nếu 1 node bị lỗi, master sẽ lên lịch lại tất cả những nhiệm vụ mà node worker đó xử lý: Các tác vụ map đã hoàn thành được thực thi lại khi bị lỗi vì đầu ra của chúng được lưu trữ trên (các) đĩa cục bộ của máy bị lỗi và do đó không thể truy cập được. Các tác vụ reduce đã hoàn thành không cần phải thực hiện lại vì đầu ra được lưu trữ trong một hệ thống tệp toàn cục.

- Khi tác vụ map được thực thi trước bởi công nhân A và sau đó sau đó được thực thi bởi công nhân B (vì A không thành công), tất cả công nhân đang thực hiện các tác vụ reduce được thông báo về việc thực hiện lại. Bất kỳ tác vụ giảm nào chưa đọc dữ liệu từ công nhân A sẽ đọc dữ liệu từ công nhân B.

* Master Failure

- Toàn bộ công việc MapReduce được khởi động lại thông qua chủ khác.

1. What is the meaning and the implication of locality? How is it used?

* Dữ liệu đầu vào được quản lý bởi GFS.

- Chọn nhóm máy MapReduce sao cho các máy đó chứa dữ liệu đầu vào trên đĩa cục bộ của chúng.

- Khi chạy các hoạt động MapReduce lớn trên một phần đáng kể công nhân trong một cụm, hầu hết dữ liệu đầu vào được đọc cục bộ và không tiêu tốn băng thông mạng.

1. Which problem is addressed by introducing a combiner function to the MapReduce model?

* Combiner được thực thi trên mỗi máy thực hiện nhiệm vụ bản đồ.

Sự khác biệt duy nhất giữa reduce function và combiner function là cách thư viện MapReduce xử lý kết quả đầu ra của hàm. Đầu ra của một reduce function được ghi vào tệp đầu ra cuối cùng. Đầu ra của một combiner function được ghi vào một tệp trung gian sẽ được gửi đến một tác vụ thu gọn. Việc kết hợp từng phần sẽ tăng tốc đáng kể các lớp hoạt động MapReduce nhất định. Phụ lục A chứa một ví dụ sử dụng bộ kết hợp.

**2.Task2 FriendCount**

Tạo file FriendCount.java

Graphical user interface, text, application

Description automatically generated

Thông dịch file FriendCount.java

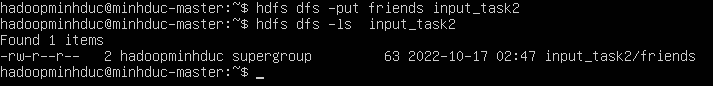
Text

Description automatically generated

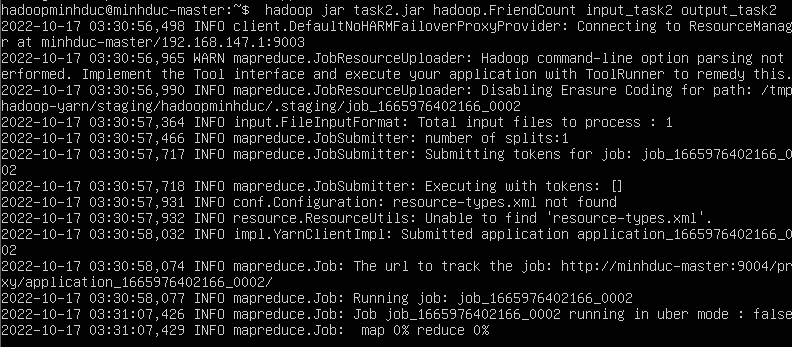
Tạo thư mực input\_task2



Đưa sample friends vào input\_tas2



Thực thi chương trình



Xuất kết quả

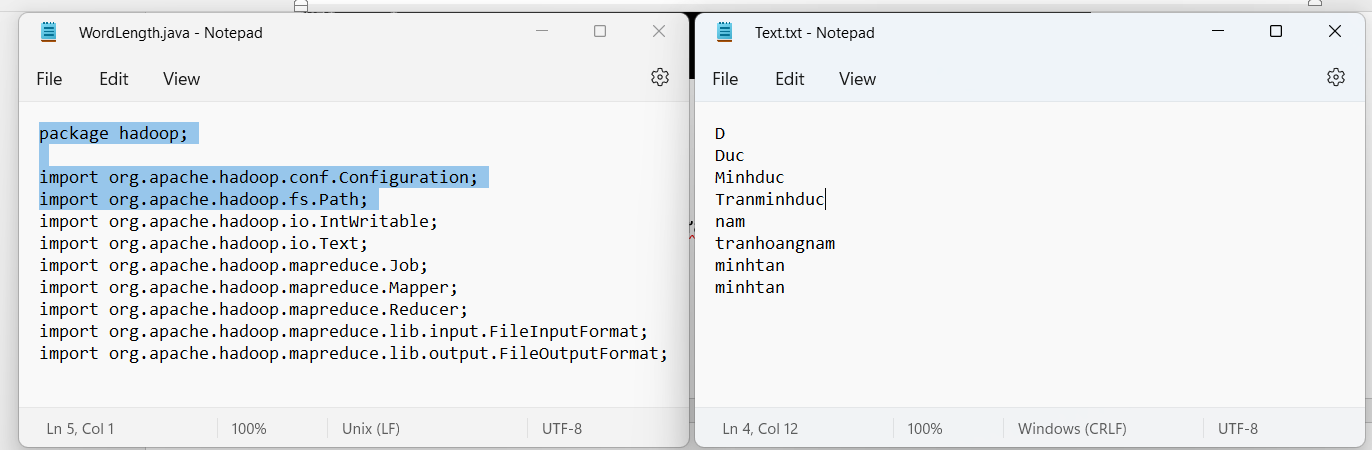
Kết quả ra là số bạn bè của một tên nào đó

Text

Description automatically generated with low confidence

**3.Task3 WordLength**

Tạo file WordLength.java và Text.txt chứa input



Thông dich chương trình

Javac -classpath hadoopcore-1.2.1.jar -d task3 WordLength.java

hadoop-core-1.2.1.jar chứa các class file cần thiết để thông dịch chương trình, -d xác định thư mục chứa các class file được tạo ra

Text

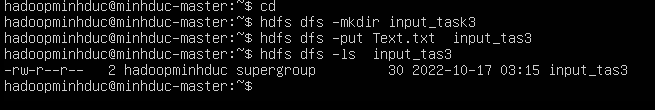
Description automatically generated

Tạo file task3.jar chứa tất cả (dấu .) file trong thư mục tassk3 (-C là lấy trong đường dẫn hiện tại)

Text

Description automatically generated

Tạo thư mục input task3(em ghi nhầm tas3 ạ)



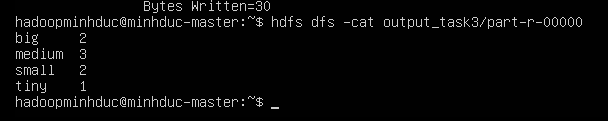
Thực thi chương trình



Text

Description automatically generated

Xuất kết quả



**4.Task4 Comprehension**

Tạo file Comprehension.java và file01 file02 chứa text

Graphical user interface, text, application

Description automatically generated

Tạo thư mục task4

Text

Description automatically generated

Thông dịch chương trình

Text

Description automatically generated

Tạo thư mục input\_task4 và đưa file file01 và file02 vào

A screenshot of a computer

Description automatically generated with medium confidence

Thực thi chương trình

Text

Description automatically generated

Xuất kết quả

Text

Description automatically generated

**5. Tài liệu tham khảo**

-<https://research.google.com/archive/mapreduce-osdi04-slides/index.html>